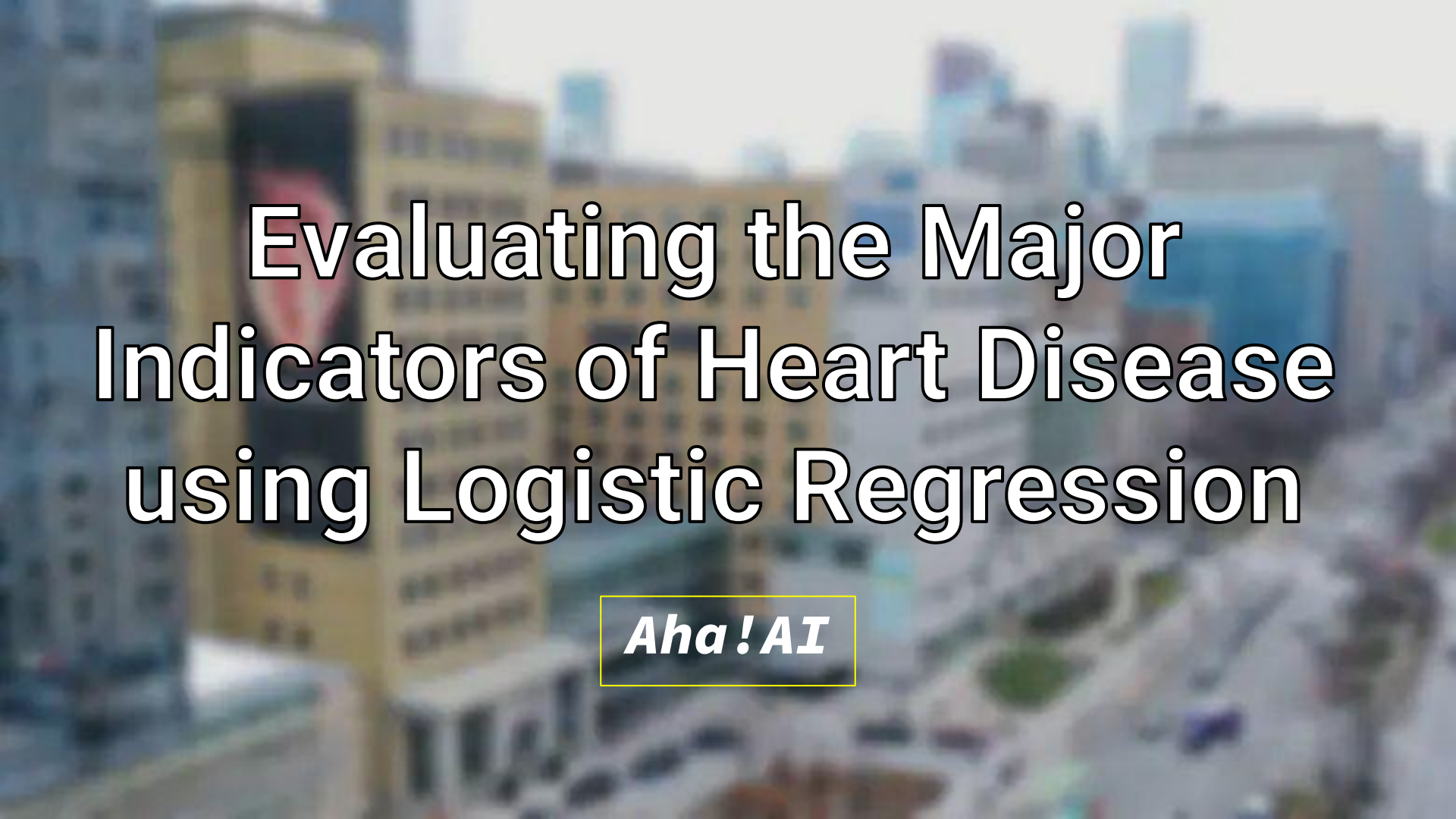
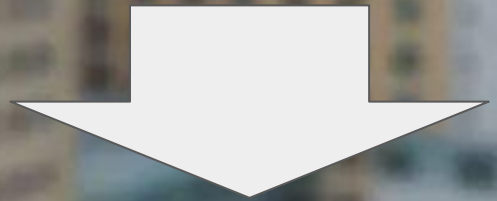
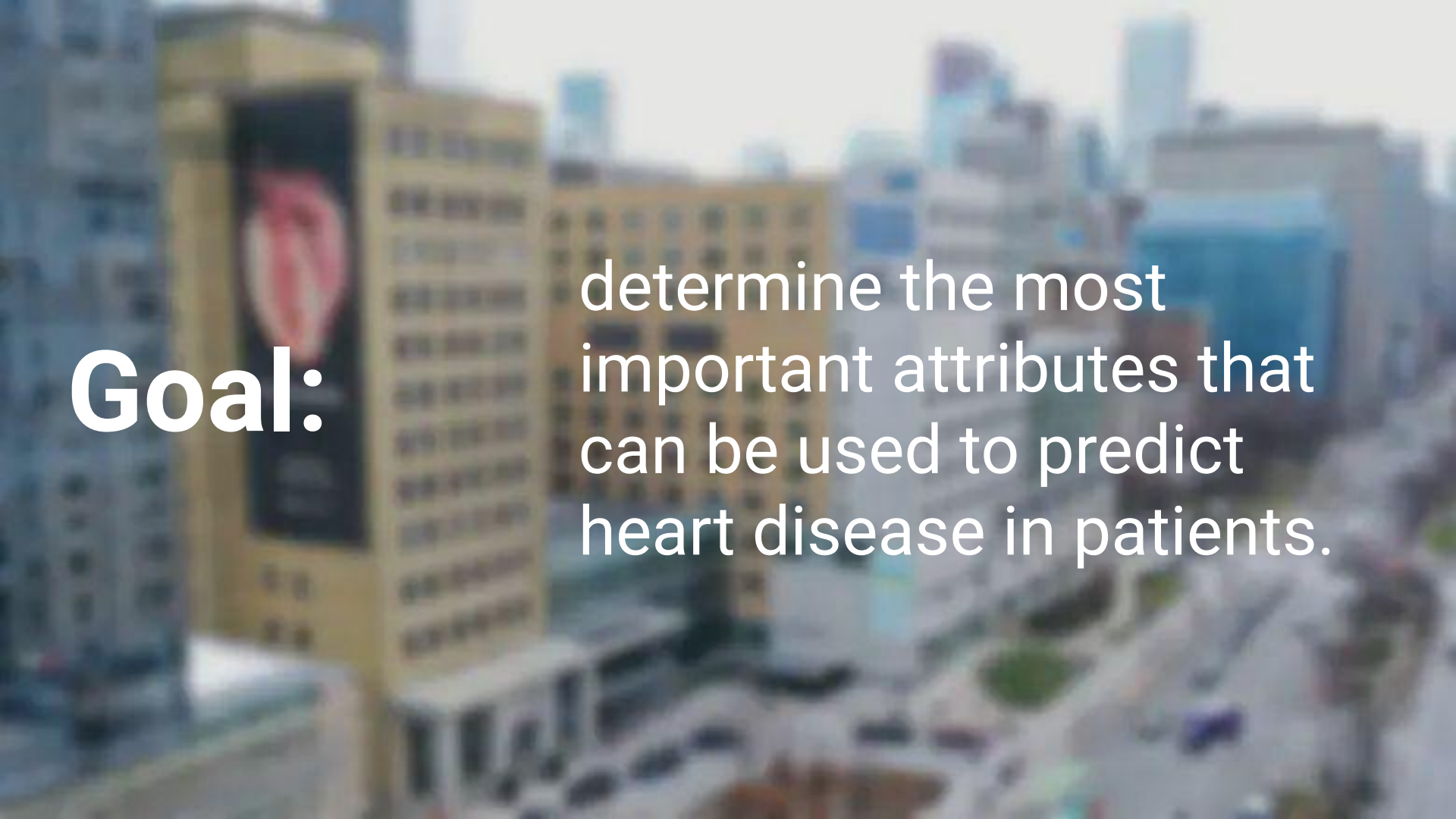

Evaluating the Major Indicators of Heart Disease using Logistic Regression
Aha!AI

Motivation

- Medical institutions have a lot of patient data
- Ample data makes it challenging to:
  - deliver meaningful patient treatment;
  - Accurately predict the likely of developing heart disease
- Almost every paper analyzed only 14 out out 76 parameters



Is this accurate? Could the other parameters be used to improve the accuracy of the model?



Goal:

determine the most important attributes that can be used to predict heart disease in patients.

Process

Collecting and parsing data

- Combined data sets of four clinics
- Assessed 66 attributes belonging to nearly 1000 patients
- Removed irrelevant information (such as patient ID) and clarified “NaN” points

Fitting data to a **logistic regression**

- Combined data sets of four clinics
- Divided the complete dataset into 75% training data, 25% testing data
- Used supervised learning in a classification problem context
- This model can predict the likelihood of an event or class

Identifying the most and least **influential attributes**

- Combined data sets of four clinics
- Trained the model for
  - The degree to which the disease is present (0-4)
  - binary classification (disease or no disease)

Performance

To what categorical degree is a disease present? *0, 1, 2, 3, 4*

```
baseline_score = logreg.score(x_test, y_test)
print(baseline_score)
```

68% Accuracy (5 Categories)

91.6% Accuracy (Collapsed to 2 Categories)

Disease or no disease? *0, 1*

```
print(np.sum((y_pred>0.5).astype(int).reshape(len(y_pred),1) == (y_test>0.5).astype(int))/len(y_test))
```

91.6% Accuracy (Trained with 2 Categories)

Results

In literature, 14 Factors are Usually Analyzed:

Age, Sex, Cp, Trestbps, Chol, Fbs, Restecg, Thalach, Exang, Oldpeak, Slope, Ca, Thal

vs. Heart Disease Level

Most Important factors we found:

Biggest Impact once removed

1. lmt
2. laddist
3. rcaprox
4. cxmain
5. ladprox

Largest Accuracy Alone

1. ramus
2. ladprox
3. rcaprox
4. om2
5. diag

Least useful factors we found:

Algorithm Improved once removed

1. painexer
2. painloc
3. slope

Least Accuracy Alone

1. thaldur
2. age
3. thalrest